CSE 598: Action and Perception

Homework 2

Yu-I Chang, yuic, 55241226

September 22, 2025

1 MAIA Implementation:

I followed the MAIA README on Great Lakes to build the MAIA environment. I initially installed under `/home`, but the 80 GB quota wasn't enough, so I switched to the 10 TB class directory. During installation I ran into some version conflicts, which I fixed by changing `open3d==0.19.0` to `0.18.0` in `requirements.txt` and adding `numpy==1.24.0` in `torch_requirements.txt`.

Next, I wired up credentials and models: I put my Hugging Face token and Gemini API key in `.bashrc`; then replaced the original `call_agent.py` with the class-provided version and updated `main.py` to use `gemini-2.5-flash`. Also, in the `sbatch` script, make sure to select either `GPU MIG40` or `SPGPU` for execution on Great Lakes since the standard GPU node does not provide enough memory. Be sure to include `conda activate maia` before running MAIA. The executable commands are in the `sbatch` file; per-model notes follow.

- **ResNet.**
  ```
  python main.py --model resnet152 --units layer4=122
  ```
  Ran successfully on the first try.

- **DINO.**
  ```
  python main.py --model dino_vits8 --units blocks.4.mlp.fc1=50
  ```
  An `ImportError` occurred during execution: `cannot import name 'trunc_normal_' from 'utils' (unknown location)`. Following a related DINO issue, I resolved it by renaming the repository's `utils/` directory to `maia_utils/` and updating all related imports to avoid the clash. Also, the exemplar data per unit is typically $\sim$100, so assign a unit index within that range.

- **CLIP.**
  ```
  python main.py --model clip-RN50 --units layer4=122
  ```
  In CLIP exemplars, the layer is labeled `visual.layer4`. During runs, this triggers a `LookupError: visual.layer4`. Although this prevented further experiments with image generation and editing, the exemplar signal remains strong. Renaming the layer in the files from `visual.layer4` to `layer4` resolves the error and results appear correct.

2 Result Validation on ResNet:

MAIA started by running 15 sample images (Figure 1) through ResNet-152 and recording the activation scores and feature maps for Unit 122 in Layer 4. These outputs were then sent to Gemini, which produced an initial hypothesis and prompts describing what the neuron might be detecting. The prompts were used with a diffusion model to generate and edit new images, which were fed back into ResNet for activation scores and feature maps, and then re-analyzed by Gemini to update the hypothesis and prompts. This cycle of "hypothesis → generation → evaluation" was repeated multiple times until the results consistently pointed to the same idea.

At first, the neuron reacted strongly to bow ties but also showed some response to neckties and other forms of formal neckwear. As more images were generated and tested, it became clear that shirts without ties or with scarves produced very low activation, while neckties gave only medium responses. After four rounds of testing, the neuron proved to be highly selective for bow ties as neckwear, mainly reacting to the distinct "bow" shape. The activation was especially strong when the bow tie appeared in the context of neckwear. By the seventh run, the interpretation stabilized, and MAIA finalized the explanation, producing the description and label shown in Figure 2.

Figure 1: Exemplar images for Unit 122 of Layer 4 in ResNet-152.

```
[DESCRIPTION]: The neuron is highly selective for the distinct visual shape of a bow tie. This includes the
characteristic "bow" knot and wing structure. The neuron shows strong activation for bow ties as standalone
objects, in illustrations, or even when integrated into other forms like pasta. Activation is consistently
maximized when the bow tie is presented as neckwear, whether worn by a person or an animal. The neuron does
not show significant activation for other types of bows or other forms of neckwear.

[LABEL]: The distinct visual shape of a bow tie, particularly when worn as neckwear.
```

Figure 2: Description from MAIA for Unit 122 of Layer 4 in ResNet-152.

3 Extension to DINO:

The reasoning process is mostly the same as with ResNet, but this time the backbone is DINO-ViT-S8, which provides the activation scores and feature maps. For this part, I chose Unit 50 in Block 4 as the exemplar. From the start, the example images (Figure 3) showed that this neuron reacts strongly to striped, checkered, and other detailed patterns. With related prompts, the diffusion model generated new images, where intricate, curvy, high-contrast designs like fingerprints produced strong activations. Barcodes and basket weaves also gave moderate responses, suggesting a preference for dense, repetitive geometric patterns. In contrast, simple patterns like wooden fences barely activated the neuron.

When the fingerprint was changed into straight parallel lines, the activation stayed high but dropped slightly. Zebra stripe edits, however, led to weaker activations, which could mean the initial zebra image didn't capture the neuron's ideal features, or that the edits removed too many key details. After six runs, it's clear that this neuron is highly selective for intricate, high-contrast patterns made of fine, closely spaced curving lines or ridges. It responds most strongly to fingerprints, spirals, and natural stripes with both high frequency and curvature, while still showing some activation for other dense patterns like parallel lines or basket weaves, as seen in the description and label in Figure 4.

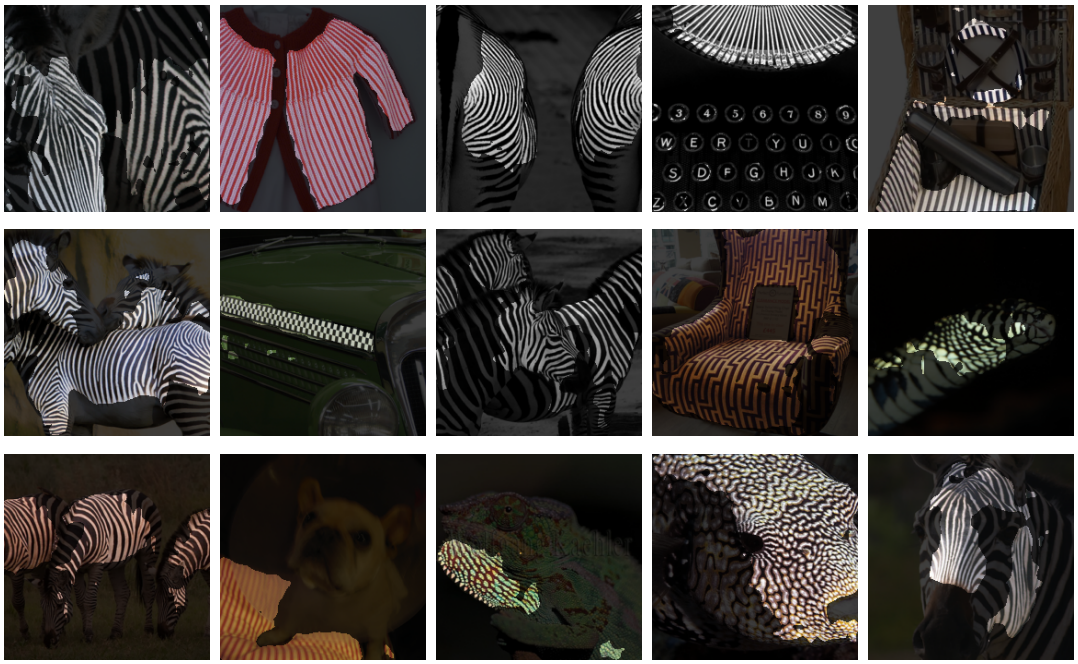

Figure 3: Exemplar images for Unit 50 of Block 4 in DINO-ViTS8.

```
[DESCRIPTION]: The neuron is highly selective for intricate, high-contrast patterns composed of closely spaced, fine
curvilinear lines or ridges. This includes patterns like fingerprints, spiraling designs, and the natural stripes found
on animals such as zebras, where the lines exhibit both high spatial frequency and significant curvature. The neuron
shows less, but still some, activation for other dense, high-contrast linear or geometric patterns (like fine parallel
lines or basket weaves), but strongly prefers those with intricate, non-rectilinear structures and fine detail.

[LABEL]: Intricate high-contrast curvilinear patterns
```

Figure 4: Description from MAIA for Unit 50 of Block 4 in DINO-ViTS8.

4 Extension to CLIP:

For this part, I switched the backbone to the CLIP-RN50 network and selected Unit 122 in Layer 4 as the exemplar (Figure 5). At first, the dataset exemplars suggested a wide range of concepts, going beyond the initial idea of just green plants or landscapes. This neuron activates strongly to groups of animals or objects, such as birds, penguins, sled dogs, and also for repeating textures and patterns like soldiers in formation, matching uniforms, pinwheels, hay bales, and dam structures. The color and specific type of object vary, suggesting the neuron is more tuned to arrangement or surface repetition than to semantic category or color.

When testing with more newly generated images, the neuron responded strongly to "dense" aggregations of distinct repeating elements, showing high activation for a large flock of penguins and moderate activation for a swarm of bees. In contrast, crowded stadiums or stacks of books gave low responses, likely because individual elements were less distinct. Similarly, textures like moss and leaves also showed weak activation, indicating that not all dense patterns are equally effective. After seven runs, the results consistently show that this neuron is selective for scenes containing many visually distinct, similar entities arranged in repetitive, ordered formations, as shown in Figure 6. Breaking that distinctness or order significantly lowers activation.

Figure 5: Exemplar images for Unit 122 of Layer 4 in CLIP-RN50.

```
[DESCRIPTION]: The neuron activates for dense aggregations of numerous, visually distinct, and similar-looking
individual entities that are arranged in prominent, repetitive patterns, especially linear rows or grid formations. This
includes concepts such as soldiers in formation, factory machines in rows, flocks of birds or animals, and fields of
uniform objects like gravestones or pinwheels.

[LABEL]: Dense, linear arrangements of distinct, similar entities
```

Figure 6: Description from MAIA for Unit 122 of Layer 4 in CLIP-RN50.